# Towards a Hybrid Audio Coder

Laurent Daudet, Stéphane Molla, Bruno Torrésani

# TOWARDS A HYBRID AUDIO CODER

L. DAUDET*
*Laboratoire d'Acoustique Musicale,*
*Université Pierre et Marie Curie,*
*11 rue de Lourmel, 75015 Paris, France*
*Email: daudet@lam.jussieu.fr*

S. MOLLA, B. TORRESANI†
*Laboratoire d'Analyse, Topologie et Probabilités,*
*Centre de Mathématiques et d'Informatique,*
*Université de Provence*
*31 rue Joliot-Curie, 13453 Marseille Cedex 13, France*
*E-mail: {molla,torresan}@cmi.univ-mrs.fr*

The main features of a novel approach for audio signal encoding are described. The approach combines non-linear transform coding and structured approximation techniques, together with hybrid modeling of the signal class under consideration. Essentially, several different components of the signal are estimated and transform coded using an appropriately chosen orthonormal basis. Different models and estimation procedures are discussed, and numerical results are provided.

## 1. Introduction

Digital audiophonic signal coding has become an important issue in many application areas. Among the most popular approaches, transform coding has received particular attention during the recent decades, as the rapid development of hardware and the discovery of novel mathematical approximation techniques has made it particularly efficient. Transform coding starts with an expansion into a suitably chosen orthonormal basis of the spaces of signals. In the (*functional*) approximation stage, only a (small) subset of the coefficients is retained, and encoded (after quantization, which is the second point where approximation comes into play).

*This work is supported by the French Ministry of Research (contract "ACI Jeunes Chercheurs").
†Work partially supported by hassip.

Although much may be done at the level of quantization, we focus here on functional approximation, which may be thought in two different ways:

- *Linear approximation:* in this approach, the retained coefficients correspond to a fixed subset of a suitably chosen basis[a].
- *Non-linear approximation:* the retained basis functions and coefficients are selected adaptively, so that the corresponding truncated expansion minimizes a given *distorsion* (some norm of the approximation error). Non-linear approximation schemes automatically outperforms the linear schemes (using the same basis and the same number of retained coefficients) in terms of distorsion, but introduces an extra costs in terms of encoding: the retained set of basis functions being not fixed *a priori*, the corresponding information (the *addresses* of retained coefficients) has to be encoded as well.

To cope with the problem of address encoding for non-linear approximation, the concept of *structured* approximation has to be introduced: it stems from the fact that for given classes of signal (and/or functional spaces), and accordingly chosen orthonormal bases, the significant coefficients have a natural tendancy to *cluster* around some given types of structures (lines/tubes, trees,...) in their index space. Exploiting such information yields substantial gains in the addresses coding, and thus in the performances of encoders. In addition, this often improves the quality of signal modeling, as we shall see on more specific examples below.

The coding schemes we shall outline in the present paper are based upon structured non-linear approximations, with an additional non-linearity, introduced for the following reason. Audio signals (like other signal classes, images[13],...) may be thought as *compound objects*, containing significantly different features[16]: mainly *tonals* (usually termed *partials* in the audio and speech literature), *transients*, and additional *residual* components. The adequate orthonormal bases for transform coding these components (at least the first two) are significantly different, and we therefore model the signal as a linear sum of two different components, each being transform coded with an adequate basis. To our knowledge, our approach is fundamentally new, in that it does not rely on a prior segmentation of the signal into different components[11,18]: superimposition is prefered to segmentation. Preliminary

[a]One generally uses an orthonormal basis which approximates well the Karhunen-Loève basis of the signal (modeled as a random signal), under some additional constraints (essentially, the existence of efficient algorithms)

results validating this approach have already been presented in previous publications[5]. We shall discuss here different levels of hybrid modeling, outline important problems (theoretical and practical) encountered in this context, and present estimation schemes developed so far.

## 2. Approximations and structures

The significantly different components present in audiophonic signals may generally be parsimoniously represented using suitable bases. This remark has led several authors to propose to expand such signals on *dictionaries*, obtained as *unions* of these bases. These dictionaries generally form quite redundant systems, and raise the problem of finding the *optimal* signal expansion on such a system (among all such expansions), optimality being understood in terms of parsimony. Focusing on the particular application to audio signals, and limiting ourselves to transient and tonal features, we are naturally led to consider a generic redundant dictionary made out of two orthonormal bases[19,20] (typically a wavelet basis $\{\psi_n, n = 0, \dots N-1\}$ and an MDCT basis $\{w_m, m = 0, \dots N-1\}$, and signal expansions of the form

$$x = \sum_{\lambda \in \Lambda} \alpha_\lambda \psi_\lambda + \sum_{\delta \in \Delta} \beta_\delta w_\delta + r \ , \tag{1}$$

where $\Lambda$ and $\Delta$ are (small, this being the *sparsity* assumption) subsets of the index sets, hereafter termed *significance maps*. In addition, we also introduce a residual signal $r$, which is not sparse with respect to the two considered bases (a *spread residual*), and is to be neglected or described differently.

Assuming that such an expansion has been obtained, the corresponding *tonal layer* and *transient layer* of the signal $x$ are defined by

$$x_{ton} = \sum_{\delta \in \Delta} \beta_\delta w_\delta \ , \quad x_{tran} = \sum_{\lambda \in \Lambda} \alpha_\lambda \psi_\lambda \ . \tag{2}$$

Several approaches have been proposed to perform such estimations, see for example[7,8,9] but most of these are not necessarily adapted when it comes to practical implementation in a coding perspective. On one hand, it is not clear that the corresponding algorithms are compatible with practical constraints, in terms of CPU and memory requirements. Also, models exploiting solely sparsity arguments cannot capture one of the main features of these signal classes, namely the *persistence* property: significant coefficients have a tendency to form "structured sets": lines or tubes of MDCT

coefficients for tonal signals, and trees[17] of wavelet coefficients for transient signals. We outline below three approaches of increasing complexity, that illustrate the benefits of introducing such "structural information" into the approximation techniques.

### 2.1. *Exploiting solely sparsity: N-term approximation*

We first describe briefly a first approach exploiting only sparsity arguments[1,5]. The tonal layer $x_{ton}$ in (2) is essentially obtained by thresholding of the MDCT transform (after suitable weighting,that corrects for the overall decay of MDCT coefficients[5]). The transient layer $x_{tran}$ is obtained similarly, by thresholding the wavelet transform of the nontonal signal $x_{nton} = x - x_{ton}$. The values of the thresholds, and therefore the corresponding significance maps $\Delta$ and $\Lambda$, are determined by the required bit rate. In order to allocate appropriate amount of bits to the two layers, the relative proportion may be pre-estimated for each time frame, using appropriate techniques[14]. The resulting algorithm reads as follows:

ALGORITHM 1: Within each time frame:
(1) Pre-estimate the relative importance of the tonal and transient layers, and corresponding bit rates.
(2) Expand the signal on an MDCT basis, and pick the largest coefficients (as estimated in Step 1). Reconstruct the *tonal signal* $x_{ton}$ and the nontonal part $x_{nton} = x - x_{ton}$.
(3) Expand $x_{nton}$ on a wavelet basis, and select the largest coefficients (according to the rule given in Step 1). This generates the transient layer $x_{tran}$, and the residual signal $x_{res} = x_{nton} - x_{tran}$.

### 2.2. *Structured N-term approximation*

As stressed in the introduction of this paper, the performances of the coding scheme are greatly improved if "structure" information is implemented into the encoder[5]. During the estimation of the tonal layer, significant MDCT coefficients are retained only if they satisfy some *time-persistence* property.

Let $\beta_\delta = \beta_{k,\nu}$ be the MDCT coefficients, where $k$ relates to time (index of the transform window $w_k$), and $\nu$ relates to frequency. Sine waves around frequency bin $\nu$ are fully characterized by the coefficients $\beta_{k\nu}$, $\beta_{k,\nu-1}$ and $\beta_{k,\nu+1}$, and the local Fourier spectrum can be well approximated[6], with a smoothed MDCT spectrum on a given window $w_t$:

$$\tilde{\beta}_{k,\nu} = \left(\beta_{k,\nu}^2 + (\beta_{k,\nu+1} - \beta_{k,\nu-1})^2\right)^{\frac{1}{2}} \tag{3}$$

Now, a width-3 tube $\mathcal{T}_\nu^{k_1,k_2} = \{\beta_{k,\nu-1}, \beta_{k,\nu}, \beta_{k,\nu+1}\}_{k=k_1 \ldots k_2}$ is retained if its averaged (over time) smoothed spectrum :

$$\sigma_p[\mathcal{T}_\nu^{k_1,k_2}] = \frac{1}{|\mathcal{T}_\nu^{k_1,k_2}|} \sum_{k=k_1..k_2} w(\nu)|\tilde{\beta}_{k,\nu}|^p \tag{4}$$

exceeds some fixed minimum value. Here, $|\mathcal{T}_\nu^{k_1,k_2}| = k_2 - k_1 + 1$ denotes the time duration of the tube, $w(\nu)$ are frequency-dependent weights (e.g. related with the absolute threshold of hearing), and $p$ is a constant selecting different types of tubes (strong short vs. long weak).

After estimation and substraction of the tonal layer, the transient layer is estimated in a similar way: the nontonal signal is expanded into a wavelet basis, and only significant wavelet coefficients satisfying some *scale persistence* property are retained. As is well known, wavelet coefficients are naturally organized into dyadic trees. A branch $\mathcal{B}$ of the wavelet coefficient tree is retained if the following modulus of regularity

$$\kappa_{q,s}[\mathcal{B}] = \frac{1}{|\mathcal{B}|} \sum_{\lambda \in \mathcal{B}} 2^{j(\lambda)s}|\alpha_\lambda|^q \ , \tag{5}$$

exceeds a given maximum value (here, $|\mathcal{B}|$ denotes the length of the branch $\mathcal{B}$, and $j(\lambda)$ denotes the scale parameter corresponding to the index $\lambda$). The constants $s, q$ characterize the considered type of transients, as they weight coefficients corresponding to different scales.

ALGORITHM 2: Within each time frame
(1) Pre-estimate the relative importance of the tonal and transient layers, and corresponding bit rates.
(2) Expand the signal on an MDCT basis, and estimate the tubes. Reconstruct the *tonal layer* $x_{ton}$ and the non-tonal part $x_{nton} = x - x_{ton}$.
(3) Expand $x_{nton}$ on a wavelet basis, and estimate the branches of the tree. Reconstruct the *transient layer* $x_{tran}$, and the residual $x_{res} = x_{nton} - x_{tran}$.

### 2.3. *Hybrid structured hidden Markov model*

The algorithms outlined above do not rely on any modeling for the considered signals, and it is therefore difficult to make any a priori estimate regarding their performances. For that purpose, a model of hybrid waveform audio model has been developed[15], which implements hidden Markov chains (resp. trees) of MDCT (resp. wavelet) coefficients.

To be more specific, a tonal signal is still modeled as in (2), and the (significant) coefficients $\beta_\delta$, $\delta \in \Delta$ are modeled as $\mathcal{N}(0, \tilde{\sigma}_\delta^2)$ independent

random variables. The index $\delta = (k, \nu)$ is a pair of time-frequency indices and the significance map $\Delta$ is characterized by a "fixed frequency" Markov chain involving two hidden states, hence by a set of initial frequencies $\nu_1, \ldots \nu_N$ and transitions matrices $\tilde{P}_1, \ldots \tilde{P}_N$ (one for each frequency bin). Therefore, we model the MDCT coefficients as a mixture model of two gaussians processes, characterizing heavy-tailed like processes. Globally, the tonal model is characterized by the set of matrices $\tilde{P}_n$, and the variances $\sigma_\delta^2$ of the two states, which are assumed to be time invariant, and on which additional constraints may be imposed.

A similar model, using Hidden Markov *trees* of wavelet coefficients[3] may be developed to describe the transient layer in the signal, as in (2). To model the *scale persistence* of large wavelet coefficients of transients, the significant wavelet coefficients $\{\alpha_\lambda, \lambda \in \Lambda\}$ of the signal are modeled as $\mathcal{N}(0, \sigma_\lambda^2)$ random variables. The index $\lambda = (j, k)$ is a pair of scale-time indices and the significance map $\Lambda$ is characterized by a "fixed time frame" Markov chain, hence by "scale to scale" transition matrices $P_j$ (with additional constraints ensuring that significant coefficients inherit a tree structure, see below.) In addition, we impose that the transition from a "non-transient" state towards a transient state is impossible, so that non significant coefficients will not be followed by significant ones. Then, a transient structure is defined within each frame as a rooted binary tree of significant coefficients.

The transient model is therefore characterized by the variances of wavelet coefficients in $\Lambda$ and $\Lambda^c$, and the persistence probability, for which estimators may be constructed. The transient states estimation itself is also performed via classical methods (involving well known Expectation-Maximization technique).

ALGORITHM 3: Within each time frame
(1) Pre-estimate the relative importance of the tonal and transient layers, and the corresponding bit rates.
(2) Learning stage: expand the signal on an MDCT basis, and estimate the best parameters for the model.
Estimation stage: estimate relevant lines.
Reconstruct the *tonal layer* $x_{ton}$ and the non-tonal part $x_{nton} = x - x_{ton}$.
(3) Learning stage: Expand $x_{nton}$ on a wavelet basis, and estimate the corresponding parameters of the Markov model.
Estimation stage: estimate transient trees.
Reconstruct the *transient layer* $x_{tran}$, and the residual $x_{res} = x_{nton} - x_{tran}$.

## 3. Numerical examples

We now illustrate the several approaches described above on numerical examples, which we use to emphasize the differences between them. For the sake of simplicity, we limit ourselves to a single sound example, namely a small piece extracted from a song of Ben Harper. The signal is exhibited at the top of the plots in FIGURES 4, 5 and 6. In all three situations, the overall number of retained coefficients (tonal + transient) was set to 5% of the number of samples of the original signal, therefore a functional compression ratio of 20. As may be seen, in all cases, the tonal layer may be seen as a copy of the original signal with a somewhat "smoothed" envelope, while the transient layer has essentially captured the "attacks". However, there are some noticeable differences between the three results, on which we elaborate below.

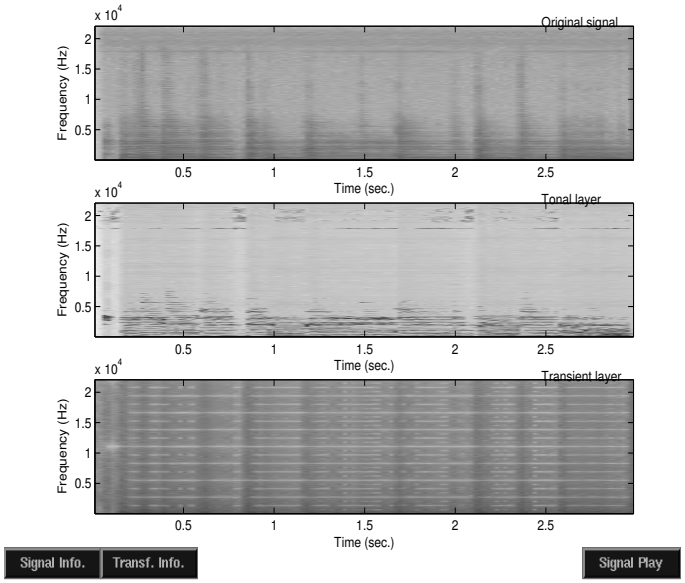


Figure 1. MDCT modulus of original signal (top), tonal (middle) and transient layers for the Ben Harper audio signal based on N-term approximation.

The main difference comes from the difference between N-term and structured N-term approximations. We have represented in Figures 4 and 5 the (logarithms of) absolute values of MDCT coefficients of the original sig-

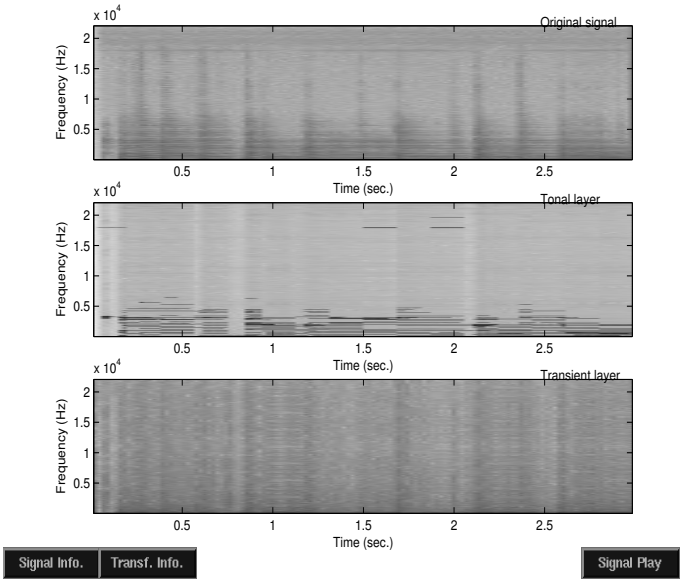


Figure 2. MDCT modulus of original signal (top), tonal (middle) and transient layers for the Ben Harper audio signal based on structured N-term approximation.

nal (top), tonal layer (middle) and transient layer (bottom). The original signal features both tonals ("horizontal structures") and transients ("vertical features"), and it may be seen than these are significantly different in Figures 4 and 5. As expected, the structured N-term approximation yields more "persistent" structures than the simple N-term approximation. This leads to a better resolution of the tonal and transient features.

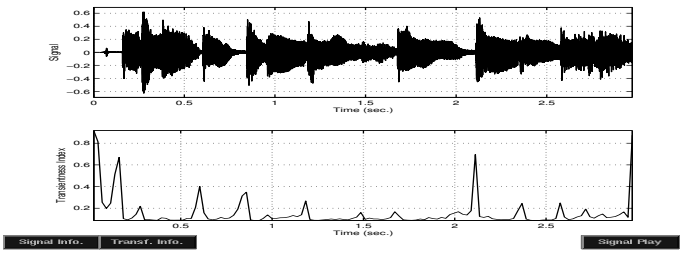


Figure 3. Transientness index for the Ben Harper signal

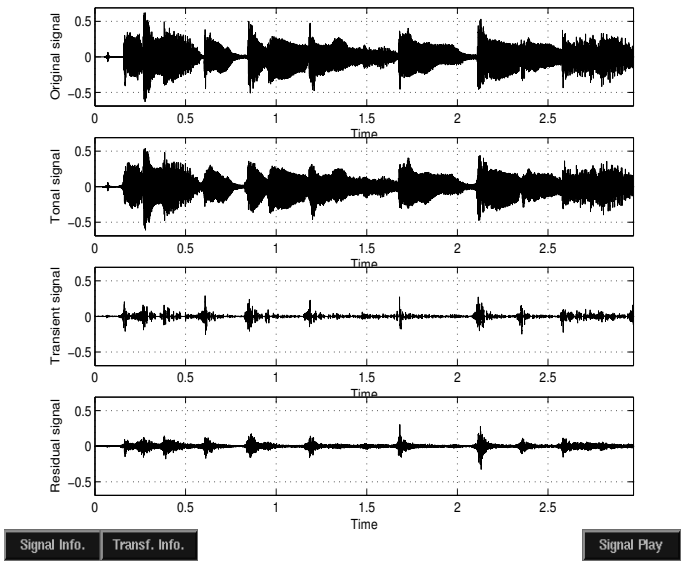


Figure 4. Hybrid expansion of the Ben Harper audio signal based on N-term approximation; from top to bottom: original, tonal layer, transient layer, residual

Concerning the signal and layers themselves, we remark that the signal at hand is mostly tonal, as confirmed by the transientness index plotted in Figure 3. Therefore, most of the energy of the signal is contained in the tonal layer, and the three methods do not make much difference regarding that particular component. Overall, between 80% and 85% of coefficients were spend on the tonal layer. While the tonal layers appear quite similar at first sight, the transient layers do exhibit significant differences. We notice that the simple N-term approximation yields transient signal present at all time, while the structured methods (structured N-term and Markov) yield more lacunary transients, which seem more relevant for the signal at hand.

The structure of the transient layer obtained in the structured N-term method appear fairly terser than in the other situations. This is presumably due to the estimation procedure, which is in that case quite simple, and does not really exploit the tree structure: let us recall that in order to keep a simple algorithm, we have limited the detection of transient structures to branches rather than complete trees. In this respect, an exploration of trees should probably be prefered at this stage, yielding higher computational costs. This point is currently under study. At the opposite, the Markov

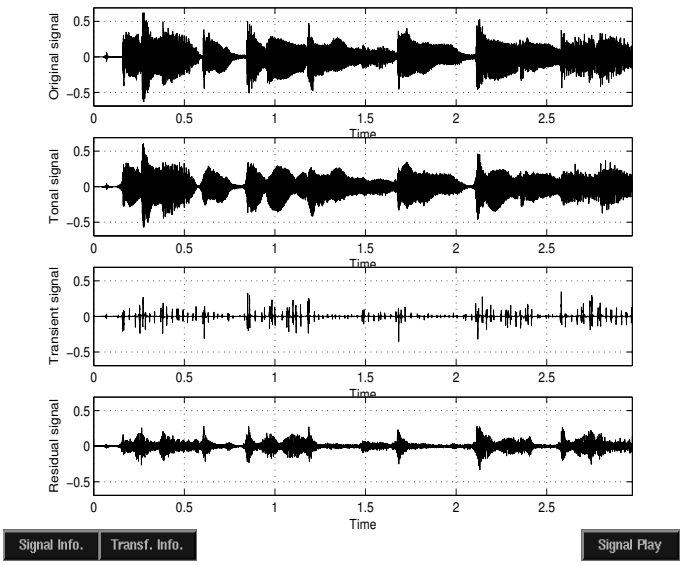


Figure 5. Hybrid expansion of the Ben Harper audio signal based on structured N-term approximation; from top to bottom: original, tonal layer, transient layer, residual

method produces a remarkably localized transient layer.

A closer examimation of the residual signal also shows that the results obtained using the Markov model are more satisfactory: indeed, the residual looks "more stationary" in that situation than in the other two cases, which is good in a coding perspective[5,15]. Again, this is probably partly due to a poor estimation of the transient layer in the N-term and structured N-term approaches.

## 4. Conclusions

We have briefly outlined in this paper three approaches of increasing complexity for hybrid encoding of audio signals. The results presented here show that hybrid modeling and structured approximation are definitely suitable techniques for audio coding. We have outlined the main differences between the three approaches, and their practical consequences. Among other conclusions, it appears clearly that a good modeling of wavelet significance trees is necessary for good estimation of the transient layer. Let us also point out that the pre-estimation of tonal and transient bit rates[14]

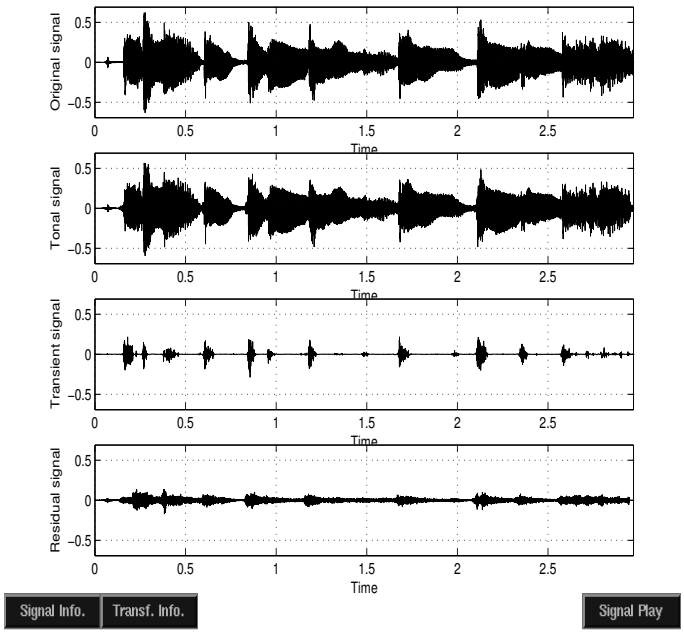


Figure 6. Hybrid expansion of the Ben Harper audio signal based on Hybrid hidden Markov model; from top to bottom: original, tonal layer, transient layer, residual

is also an important ingredient, and that our current approach probably needs further refinement.

More complete results in a signal encoding context will be presented in a forthcoming publication. Additional illustrations and sounds available at

`http://www.cmi.univ-mrs.fr/~torresan/papers/Chongqing`

## References

1. J. Berger, R. Coifman and M. Goldberg, Removing noise from music using local trigonometric bases and wavelet packets. *J. Audio Eng. Soc.*, **42**(10) (1994), pp. 808–818.
2. A. Cohen, W. Dahmen, I. Daubechies and R. DeVore, Tree approximation and optimal encoding. *Appl. Comput. Harmon. Anal.* **11**(2) (2001), pp.192–226.
3. M. S. Crouse, R. D. Nowak and R. G. Baraniuk, Wavelet-Based Signal Processing using Hidden Markov Models, *IEEE Transactions on Signal*

*Processing* **46** (1998), pp. 886–902.
4. I. Daubechies, *Ten lectures on wavelets*. SIAM, Philadelphia, PA.
5. L. Daudet and B. Torrésani, Hybrid models for audio signals encoding, *Sig. Proc.* **16** (2002), 793–810
6. L. Daudet and M. Sandler, MDCT analysis of sinusoids: exact results and applications to coding artifacts reduction, *IEEE Trans. ASSP*, to appear (2004).
7. D.L. Donoho and X. Huo, Uncertainty principles and ideal atomic decompositions, *IEEE Trans. Inf. Th.* **47**(7) (2001), 2845-2862.
8. M. Elad and A.M. Bruckstein, A generalized uncertainty principle and sparse representations, *IEEE Trans. Inf. Th.* **48**(9) (2001), 2558-2567.
9. R. Gribonval and M. Nielsen, Sparse representations in union of bases, Technical Report 1499, Institut National de Recherches en Informatique et Automatique, IRISA Rennes (2003).
10. N. S. Jayant and P. Noll. *Digital coding of waveforms*. Prentice-Hall, 1984.
11. S. Levine. *Audio Representations for Data Compression and Compressed Domain Processing*. PhD thesis, Stanford University, 1998.
12. S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, **41** (1993), 3397–3415.
13. F.G. Meyer, A.Z. Averbush and R.R. Coifman, Multilayered Image Representation: Application to Image Compression, *IEEE Transactions on Image Processing* **11** (2002), 1072–1080.
14. S. Molla and B. Torrésani, Determining local transientness of audio signals, *IEEE Signal Processing Letters*, to appear (2004).
15. S. Molla and B. Torrésani. An Hybrid Audio Scheme using Hidden Markov Models of Waveforms, Preprint, Sept. 2003, submitted to *Appl. and Comp. Harm. Anal.*
16. X. Serra. *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition*. PhD thesis, Stanford University (1989).
17. J. M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans. Signal Processing*, **41**(12) (1993), 3445–3462.
18. T. Verma, *A Perceptually Based Audio Signal Model With Application to Scalable Audio Compression*, PhD thesis, Stanford University (2000).
19. M. Vetterli and J. Kovacevic. *Wavelets and subband coding*. Prentice Hall, Englewood Cliffs, NJ, USA, 1995.
20. M. V. Wickerhauser. *Adapted Wavelet Analysis from Theory to Software*. AK Peters, Boston, MA, USA, 1994.